Natural Language Processing for Early Detection of Mental Health

Shaoxiong Ji

Technical University of Darmstadt
Email: shaoxiong.ji@tu-darmstadt.de

GeMTeX Large Language Model Workshop
18 November 2024

Warning: this presentation contains text examples that are negative, depressive, or adverse.

TECHNISCHE UNIVERSITÄT DARMSTADT

Mental Health in Social Content

- Mental health is a critical issue
- Mental disorders could sometimes turn to suicidal ideation
- Early detection from social content and early prevention

Figure: Warning signs of mental illness[a]

[a] Source https://www.nami.org/Blogs/NAMI-Blog/May-2022/Understanding-The-Early-Warning-Signs-of-Mental-Illness

TECHNISCHE UNIVERSITÄT DARMSTADT

Significance in Mental Health

- **Language in Psychotherapy:** Exploration of linguistic expression in psychotherapy reveals emotional states and contributes to psychopathological networks.
- **Psychiatric Diagnostics:** Language plays a crucial role in diagnosing and understanding mental health conditions.
- **AI's Role:** Leveraging NLP, ML, and AI techniques can lead to innovative solutions for improving mental illness diagnostics and therapy.

TECHNISCHE UNIVERSITÄT DARMSTADT

Early Detection

Data set construction[1,2]

- Social media, e.g., Reddit and Twitter
- Weak labels and manual annotations

Tasks

- suicidal ideation detection
- mental health classification (depression, anxiety, stress, and bipolar)

Dataset	SID	MH
SWMH	10,182	44,230
T-SID	594	9,694

[1] S. Ji, C. P. Yu, S.-f. Fung, S. Pan, and G. Long. "Supervised Learning for Suicidal Ideation Detection in Online User Content". In: *Complexity* (2018).

[2] S. Ji, X. Li, Z. Huang, and E. Cambria. "Suicidal Ideation and Mental Disorder Detection with Attentive Relation Networks". In: *Neural Computing and Applications* 34 (13 2022), pp. 10309–10319.

Early Detection

Table: Annotation rules and examples of social texts[3]

Categories	Rules	Examples
Suicide Text	• Expressing suicidal thoughts • Including potential suicidal actions	*I want to end my life tonight.* *Yesterday, I tried to cut my wrist, but failed.*
Non-suicide Text	• Formally discussing suicide • Referring to other's suicide	*The global suicide rate is increasing.* *I am so sad to hear that Robin Williams ended his life.*
	• Not relevant to suicide	*I love this TV show and watch every week.*

[3] S. Ji, C. P. Yu, S.-f. Fung, S. Pan, and G. Long. "Supervised Learning for Suicidal Ideation Detection in Online User Content". In: *Complexity* (2018).

Domain-specific LMs for Mental Health

MentalBERT[4]

- English posts collected from Reddit
- Continual pretraining

Model	UMD Rec.	UMD F1	T-SID Rec.	T-SID F1	SWMH Rec.	SWMH F1	SAD Rec.	SAD F1	Dreaddit Rec.	Dreaddit F1
BERT	61.63	58.01	88.44	88.51	69.78	70.46	62.77	62.72	78.46	78.26
RoBERTa	59.39	**60.26**	88.75	88.76	**70.89**	72.03	66.86	67.53	80.56	80.56
BioBERT	57.76	58.76	86.25	86.12	67.10	68.60	66.72	66.71	75.52	74.76
ClinicalBERT	58.78	58.74	85.31	85.39	67.05	68.16	62.34	61.25	76.36	76.25
MentalBERT	**64.08**	58.26	88.65	88.61	69.87	71.11	67.45	67.34	80.28	80.04
MentalRoBERTa	57.96	58.58	**88.96**	**89.01**	70.65	**72.16**	**68.61**	**68.44**	**81.82**	**81.76**

[4] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria. "MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare". In: *Proceedings of LREC*. Marseille, France: European Language Resources Association, 2022, pp. 7184–7190.

Domain-specific LMs for Mental Health

Mental Longformer[5]

- More data
- Long-range ability

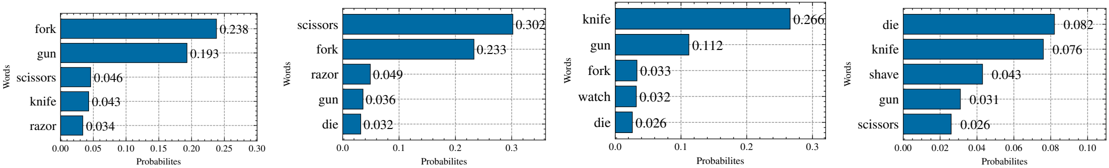
Dataset	Seq. Len.	Longformer Rec.	Longformer F1	MentalLongformer Rec.	MentalLongformer F1
UMD	512	65.10	58.36	62.24	**59.74**
	1024	63.27	62.34	**64.29**	**66.22**
	1536	67.55	66.90	67.55	66.90
	2048	65.92	67.90	**70.82**	**69.19**
	2560	68.98	68.10	**72.04**	**72.53**
	3072	71.43	72.15	**72.65**	69.76
	3584	62.65	66.08	**72.45**	**72.13**
	4096	74.29	72.85	73.06	72.47
CLP	512	64.33	63.44	59.00	54.85
	1024	70.67	69.68	**71.33**	**70.76**
	1536	69.00	67.27	**71.00**	**69.57**
	2048	75.33	75.26	**72.32**	**72.00**
	2560	75.00	74.57	**76.00**	**75.69**
	3072	65.33	62.53	**72.33**	**70.97**
	3584	72.00	70.91	**75.00**	**74.31**
	4096	75.67	75.47	**77.00**	**76.32**

[5] S. Ji, T. Zhang, K. Yang, S. Ananiadou, E. Cambria, and J. Tiedemann. "Domain-specific Continued Pretraining of Language Models for Capturing Long Context in Mental Health". In: *arXiv preprint arXiv:2304.10447* (2023).

TECHNISCHE UNIVERSITÄT DARMSTADT

User Intention

But can the pretrained models "understand" the latent intention to some extent?[6]

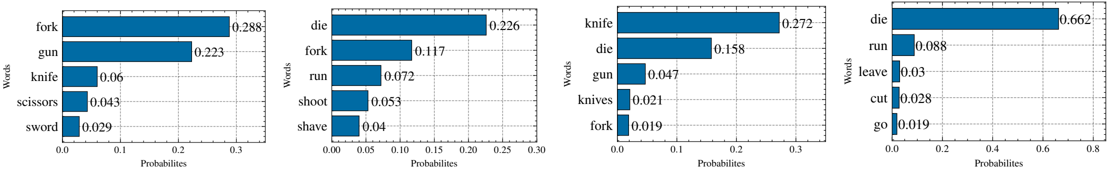


(a) BERT (b) RoBERTa (c) MentalBERT (d) MentalRoBERTa

Figure: The output of word probabilities in the fill-mask language modeling task using various pretrained masked language models for the sentence "I am going to buy a knife and [MASK]."

[6] S. Ji. "Towards Intention Understanding in Suicidal Risk Assessment with Natural Language Processing". In: *Findings of EMNLP*. Association for Computational Linguistics, 2022, pp. 4028–4038.

User Intention



(a) BERT (b) RoBERTa (c) MentalBERT (d) MentalRoBERTa

Figure: The output of word probabilities in the fill-mask language modeling task using various pretrained masked language models for the sentence "This life is not worth living. I am going to buy a knife and [MASK]."

TECHNISCHE UNIVERSITÄT DARMSTADT

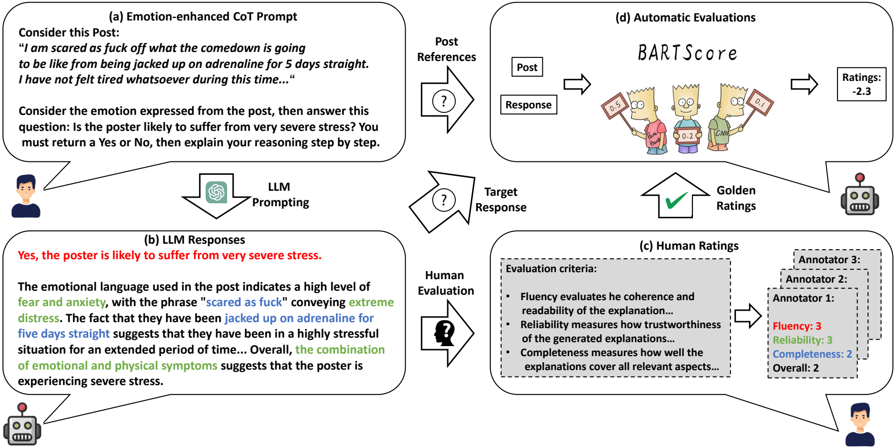
Large Language Models

LLMs, esp. ChatGPT, changed many things!

- Scalable Support: Preliminary support for individuals needing mental health resources.
- Data-Driven Insights: Analyzes text data (e.g., social media posts, clinical notes) for trends in mental health conditions, symptomatology, and public sentiment.
- Augmenting Mental Health Tools: Complements traditional screening tools with conversational interfaces, symptom monitoring, and intervention recommendations.

TECHNISCHE UNIVERSITÄT DARMSTADT

LLM for Mental Health Analysis



TECHNISCHE UNIVERSITÄT DARMSTADT

LLM for Mental Health Analysis

Emotion-enhanced prompts and evaluation of LLMs[7]

- **Emotion-enhanced CoT prompting**: emotion-enhanced zero-shot chain-of-thoughts prompts
- **Supervised emotion-enhanced prompting**: adding the sentiment/emotion labels distantly supervised by sentiment lexicons to the proper positions of the zero-shot prompt
- **Few-shot Emotion-enhanced Prompts**: domain experts write one response example for each label class within a test set

[7] K. Yang, S. Ji, T. Zhang, Q. Xie, Z. Kuang, and S. Ananiadou. "Towards Interpretable Mental Health Analysis with Large Language Models". In: *Proceedings of EMNLP*. 2023.

Instruction fine-tuned LLMs for Mental Health

MentaLLaMA[8]

1). Task-specific instruction:
[DR]: *You will be presented with a post and an assigned label to identify whether the poster shows symptoms of depression. Consider this post to explain the reasoning of the label step by step. Here are four examples:*

2). Expert-written examples:
[DR]:
Example 1:
Post: Does everyone else just hurt all the time It's not like physical pain or soreness, it's just this overwhelming feeling of exhaustion …
Response: Yes. Reasoning: The post conveys a deep sense of emotional pain, exhaustion, and numbness …
Example 2:
Post: Hello!:) I'm a new user so if this post ends up in a weird place/thread …
Response: No. Reasoning: The post does not exhibit strong emotional indicators of very severe depression …
Example 3:
…

3). Query for the target post:
[DR]:
Post: How to avoid a relapse? I've been having a particularly rough year; I attempted suicide…
Response: Yes. Reasoning:

[8] K. Yang, T. Zhang, Z. Kuang, Q. Xie, and S. Ananiadou. "MentalLLaMA: Interpretable Mental Health Analysis on Social Media with Large Language Models". In: *arXiv preprint arXiv:2309.13567* (2023).

Instruction fine-tuned LLMs for Mental Health

More instruction-tuned models:

- Mental-LLM[9]: various LLMs on mental health prediction tasks using online text data
- Psy-LLM[10]: to support psychological counseling by providing question-answering capabilities in mental health consultation settings

Instruction fine-tuning improve the predictive performance.

[9] X. Xu, B. Yao, Y. Dong, S. Gabriel, H. Yu, J. Hendler, M. Ghassemi, A. K. Dey, and D. Wang. "Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data". In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8.1 (Mar. 2024). DOI: 10.1145/3643540. URL: https://doi.org/10.1145/3643540.

[10] T. Lai, Y. Shi, Z. Du, J. Wu, K. Fu, Y. Dou, and Z. Wang. *Psy-LLM: Scaling up Global Mental Health Psychological Services with AI-based Large Language Models*. 2023. arXiv: 2307.11991 [cs.CL].

Table: MentaLLaMA Evaluation results of correctness on the IMHI test set. All results are weighted F1 scores. "Param." denotes the number of parameters for each model. In zero-shot/few-shot Methods, "ZS" denotes zero-shot methods, and "FS" denotes few-shot methods. The best values in discriminative and interpretable mental health analysis methods are highlighted in bold.

Model	Param.	CAMS	CLP	DR	Dreaddit	IRF	loneliness	MultiWD	SAD	SWMH	T-SID
Discriminative methods											
BERT-base	110M	34.92	62.75	90.90	78.26	72.30	83.92	**76.69**	62.72	70.76	88.51
RoBERTa-base	110M	36.54	66.07	95.11	80.56	71.35	83.95	–	67.53	72.03	88.76
MentalBERT	110M	39.73	62.63	**94.62**	80.04	**76.73**	82.97	76.19	67.34	71.11	88.61
MentalRoBERTa	110M	**47.62**	**69.71**	94.23	**81.76**	–	**85.33**	–	**68.44**	**72.16**	**89.01**
Zero-shot/few-shot methods											
LLaMA2-7B$_{ZS}$	7B	16.34	36.26	58.91	53.51	38.02	58.32	40.1	11.04	37.33	25.55
LLaMA2-13B$_{ZS}$	13B	14.64	39.29	54.07	36.28	38.89	55.48	53.65	13.2	40.5	25.27
ChatGPT$_{ZS}$	175B	33.85	56.31	82.41	71.79	41.33	58.40	62.72	54.05	49.32	33.30
ChatGPT$_{FS}$	175B	44.46	61.63	84.22	75.38	43.31	58.78	64.93	63.56	60.19	43.95
GPT-4$_{FS}$	1.76T	42.37	**62.0**	82.0	78.18	51.75	72.85	62.58	55.68	62.94	40.48
Completion-based fine-tuning methods											
T5-Large	770M	40.2	48.6	84.9	77.7	74.0	80.8	76.4	58.1	70.0	77.1
BART-Large	406M	43.8	50.3	84.6	**80.0**	76.2	83.3	**77.2**	59.6	71.5	**77.9**
LLaMA2-7B	7B	30.47	51.17	84.94	61.59	73.5	81.25	65.52	49.6	63.08	68.93
Instruction-tuning methods											
MentaLLaMA-7B	7B	32.52	59.86	76.14	71.65	67.53	83.52	68.44	49.93	72.51	72.64
MentaLLaMA-chat-7B	7B	44.8	51.84	83.95	62.2	72.88	83.71	75.79	62.18	**75.58**	77.74
MentaLLaMA-chat-13B	13B	**45.52**	52.61	**85.68**	75.79	**76.49**	**85.1**	75.11	**63.62**	71.7	75.31

Ethics

- Bias risks: LLMs can reflect gender and racial biases, potentially reinforcing mental health disparities.
- Sources of bias: Inherited from human-generated content and labeling practices, including stereotypes and confirmation bias.
- Deployability gaps: Ethical concerns impact all stages of use

Rethinking LLMs in Mental Health Applications

- LLM-generated explanation $\neq$ interpretablity
- LLMs as a user interface

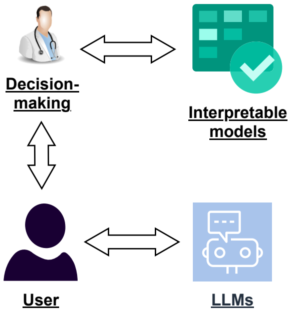


Figure: LLMs as a user interface[11]

[11] S. Ji, T. Zhang, K. Yang, S. Ananiadou, and E. Cambria. "Rethinking Large Language Models in Mental Health Applications". In: *arXiv preprint arXiv:2311.11267* (2023).

Mental Health in Multilingal Scenarios

Mostly English benchmarks

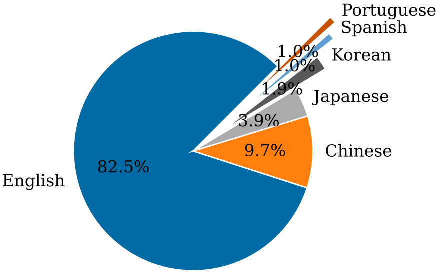


Figure: The availability of multilingual texts for mental health research in 2021[12]

[12] K. Harrigian, C. Aguirre, and M. Dredze. "On the State of Social Media Data for Mental Health Research". In: *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*. ACL, 2021, pp. 15–24.

Mental Healthcare Goes Multilingal

- A collection of multilingual social text [13]
  - social media
  - Wikipedia
  - mental health forums
  - mental health-related textbooks
- Manual translation from English to other languages for cross-lingual evaluation

[13] Work in progress

Conversational AI

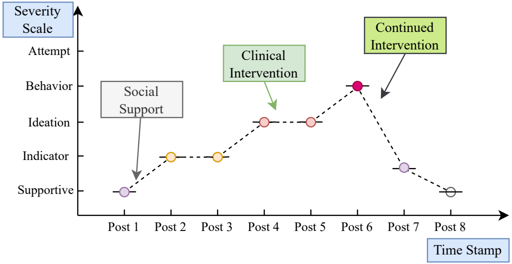
In mental health counseling[14]

- the nuances of individual experiences - individual training data and personalized responses
- empathetic and contextual understanding that human counselors possess
- helpfulness and harmfulness

[14] S. Ji, T. Zhang, K. Yang, S. Ananiadou, and E. Cambria. "Rethinking Large Language Models in Mental Health Applications". In: *arXiv preprint arXiv:2311.11267* (2023).

Post History / Therapist-Patient Conversations

Sequential risk assessment[15]



Language analysis of therapist-patient conversations

- Contextual understanding
- Memory states
- Retrieval augmentation

[15] S. Ji. "Towards Intention Understanding in Suicidal Risk Assessment with Natural Language Processing". In: *Findings of EMNLP*. Association for Computational Linguistics, 2022, pp. 4028–4038.

Takeaways

- Continual pretraining improves prediction
  - but stuck in intention "understanding", esp., for small models
- Decoder-only LLMs still behind encoder-only small models;
  - but can handle mutliple tasks with one model
- Several challenges in NLP and LLMs for mental health
  - interpretability;
  - multilingual and multicultural settings;
  - conversational understanding